ESUKA – JEFUL 2013, 4 – 1: 185 – 204

KOLLOSTRUKTSIOONILISED MEETODID JA KONSTRUKTSIOONILISE VARIEERUMISE TUVASTAMINE

Kristel Uiboaed
Tartu Ülikool

Kokkuvõte. Käesolev artikkel uurib kahe verbi ühendeid ja nende varieerumist eesti murretes. Artikkel tutvustab üht kvantitatiivse korpuslingvistika meetodit – kollostruktsioonilist analüüsi – ja rakendab seda eesti murrete verbikonstruktsioonide tuvastamiseks. Kui kõigi murrete verbikonstruktsioonid on tuvastatud, vaadeldakse nende varieerumist murretes ning esitatakse eesti murrete jaotus verbikonstruktsioonide alusel. Tulemused osutavad, et verbikonstruktsioonipõhine varieerumine murretes viib hoopis teistsuguse rühmituseni kui traditsiooniline murdejaotus. Verbikonstruktsioonidevahelised erinevused murretes on kõige suuremad ida- ja läänepoolsemate murrete vahel, kusjuures idapoolsemad murded kasutavad eri konstruktsioone märkimisväärselt vähem kui läänemurded.

Märksõnad: ahelverbid, dialektoloogia, kollostruktsioonlised meetodid, korpuslingvistika, murdesüntaks, verbikonstruktsioonid

1. Sissejuhatus

Järgnev artikkel annab ülevaate ühest statistilisest keelekonstruktsioonide analüüsi meetodist – kollostruktsioonilisest analüüsist. Kollostruktsioonilise analüüsi meetodid on kvantitatiivse korpuslingvistika meetodite perekond, mille on välja töötanud A. Stefanowitsch ja S. Gries (2003). Kollostruktsioon on sulam sõnadest kollokatsioon ja konstruktsioon. Metodoloogia on edasiarendus rohkem tuntud kollokatsioonide tuvastamise meetoditest (Evert 2005; 2008). Kui kollokatsioonide ehk püsiühendite tuvastamise meetodid on huvitatud ainult kahe või enama sõna koosesinemissagedustest ja nende statistilisest seosest, keskendumata oluliselt muudele süntaktilistele või semantilistele teguritele, siis kollostruktsiooniline analüüs keskendub

just keelekonstruktsioonidele ning nendega seotud sõnade omavahelistele suhetele. Analüüsimeetod on omaks võtnud terminoloogia ning analüüsieeldused konstruktsioonigrammatika raamistikust (Stefanowitsch, Gries 2009), mille järgi peetakse keelekonstruktsiooniks kindla keelelise funktsiooniga vormi ja tähenduse ühisusi (nt Goldberg 1995)[1].

Järgnevalt tutvustan kollostruktsioonide analüüsi meetodeid eesti murrete korpuse näitel, st uurib konstruktsioonilist varieerumist eesti murretes. Käesolevas artiklis kasutan kollostruktsioonilisi meetodeid verbikonstruktsioonide tuvastamiseks eesti murretes ning saadud tulemusi murrete grupeerimiseks. Meetod võimaldab uurida konstruktsioone väga erinevatest aspektidest, näiteks konstruktsioonide grammatiseerumisprotsessi (Hilpert 2006, 2008), sünonüümsete konstruktsioonide detailsemaid erinevusi (Gilquin 2006), konstruktsioonilist varieerumist (Mukherjee, Gries 2009) jne.

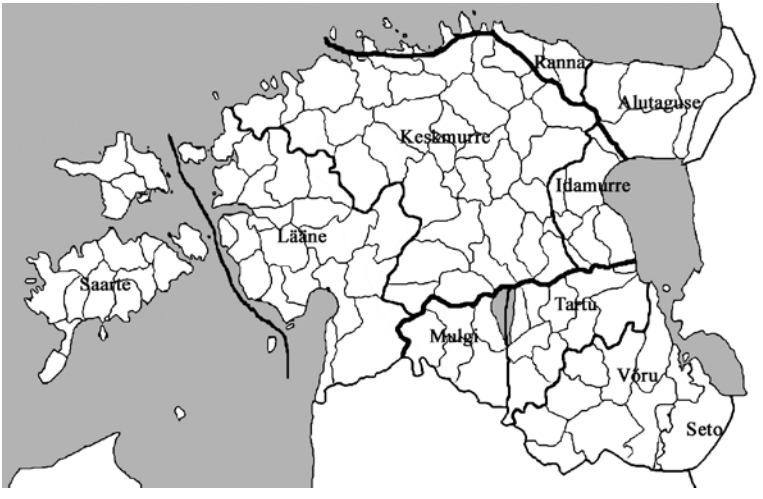
Kui konstruktsioonid on tuvastatud ning nende statistilise tugevuse väärtused eri murretes välja arvutatud, jätkan saadud info põhjal murrete grupeerimisega. Murrete ning nende verbikonstruktsioonide grupeerimiseks ja tulemuste visualiseerimiseks kasutan üht mitmemõõtmelise statistika meetodit – korrespondentsanalüüsi.

Artikkel on üles ehitatud järgnevalt. Teises osas annan ülevaate kasutatud materjalist ning selle kogumise viisist. Kolmandas osas tutvustan kollostruktsioonilise analüüsi mõistmiseks vajalikku terminoloogiat. Neljas osa annab lühikese ülevaate sõnadevahelise seose tugevuse mõõtmise statistikutest ning viies osa tutvustab kollostruktsioonlisi meetodeid. Kuuendas osas esitan tehtud katse tulemused.

1 Eestikeelset ülevaadet konstruktsioonigrammatikast vt Penjam (2008) ja Sahkai (2011)

2. Materjal

Iga korpuspõhine uurimistöö algab automaatse või poolautomaatse materjali kogumisega. Kuna käesolev artikkel tegeleb konstruktsioonilise varieerumisega eesti murretes, tuleb materjal eesti murrete korpusest[2]. Eesti murrete korpus on elektrooniline andmebaas, mis sisaldab tekste kõigist eesti murretest. Korpuses kasutatav murdejaotus on esitatud joonisel 1. Kõik korpuse tekstid on morfoloogiliselt märgendatud (üle 600 000 sõna), st igale sõnale on lisatud sõnaliigi jm morfoloogiline informatsioon (murdekorpuse märgenduse kohta täpsemalt vt Lindström jt 2006). See teeb teksti automaatselt kergesti töödeldavaks.



Joonis 1. Eesti murrete jaotus murdekorpuses[3].

Käesolev artikkel tegeleb finiitse ja infiniitse verbivormi ühenditega ning püüab statistiliste meetoditega tuvastada, milliseid sellistest ühenditest võiks pidada konstruktsioonideks konstruktsioonigrammatika mõttes. Enne kui analüüsimine või-

2 Eesti murrete korpus. www.murre.ut.ee [01.09.2010]
3 Alutaguse = kirdemurre.

malikuks saab, on vajalik kõigi potentsiaalsete konstruktsioonide tuvastamine murdekorpusest. Verbi finiitne ja infiniitne vorm moodustavad EKG järgi ahelverbe, mille semantilist sisu kannab käändeline verbivorm ning pöördeline vorm modifitseerib käändelise vormiga esitatavat situatsiooni, lisades kogu konstruktsioonile kausatiivse, viisi, modaalsuse või aspekti tähenduse (Erelt jt 1993: 19–20). Paljud sellistes konstruktsioonides esinevad finiitsed verbid on tugevasti grammatiseerunud, kui nad esinevad koos teatud infiniitsete vormidega (*siis pere **pidi** riidu **minema,** elu **tahtis elädä***), seetõttu käsitleb käesolev artikkel ainult finiitse verbi lemma ja infiniitse verbi morfoloogilise kategooria paare (V_{fin} + inf). Artikli eesmärk on seega välja selgitada, millised V_{fin} + inf paarid võivad olla teatud määral grammatiseerunud.

Kuna kaks uuritavat verbivormi saavad konstruktsiooni moodustada vaid siis, kui nad esinevad koos samas osalauses, on oluline, et potentsiaalsed konstruktsioonid ei ületaks osalause piire. Selleks oli vajalik tähistada osalausete piirid. Osalausestamist oli võimalik teha automaatselt eesti keele süntaksianalüsaatoriga (Müürisep 2000), mida on kohandatud murrete automaatanalüüsi tarbeks (Lindström, Müürisep 2009). Tähistatud osalausepiirid võimaldasid moodustada V_{fin} + inf paare automaatselt osalausete piire ületamata.

Analüüsiks vajalik andmestik on koostatud järgnevalt:

1) Loeti kokku ainult infiniitse verbi morfoloogilise kategooria märgendid; selles vormis esinev verb ei olnud oluline, näiteks mitu *da*-infinitiivi, *ma*-infinitiivi jne igas murdes esines.
2) Kokku loeti kõik finiitsed verbid, mis sellistes paarides esinesid. Materjali kaasati ainult verbide algvormid, arvestamata nende aega, tegumoodi jmt.
3) Kogu V_{fin} + inf paari sagedus saadi seega infiniitse verbivormi morfoloogilise kategooria ja temaga samas osalauses esineva finiitse verbi esinemisarvudest. Näiteks *jõuame joosta* ja *jõudis seletada* on loetud samaks ühendiks (*da*-infinitiiv + *jõudma*), kuna lugesin kokku vaid finiitse verbi (*jõudma*) ja infiniitse verbivormi kategooria

(*da*-infinitiiv); infiniitne verb ise pole oluline. Ühendid liigitati eraldi vaid infiniitse verbivormi alusel, näiteks *jõudis joosta* ja *jõudis sööma* on erinevad ühendid (*da*-infinitiiv + *jõudma* ja *ma*-infinitiiv + *jõudma*), sest nende ühendite infiniitne osis on erinev. Potentsiaalse konstruktsiooni tekstisagedus on seega arv, mitu korda see ühend (infiniitne vorm ja finiitne verb) samas osalauses koos esines.

Ükski selline automaatne keeletöötlusprotsess ei ole ideaalne, st teatud vigade hulk on vältimatu. Siiski on võimalik arvutada, kui täpne selline verbikonstruktsioonide kandidaatide tuvastamisprotsess on. Täpsuse arvutamiseks moodustasin 5000-sõnalise testkorpuse, kuhu valisin juhuslikult igast murdest 500 sõna. Sealt oli näha, et sellise ekstraheerimisprotsessi täpsus oli 80%. Peale madala sagedusega kombinatsioonide (<3) eemaldamist tõusis täpsus 92%-ni. Eemaldasin korpuses vähem kui kolm korda esinevad paarid ka oma lõplikust analüüsist.

Eesti keele sõnajärg on suhteliselt varieeruv. Pöördeline verb esineb pealauses sagedamini teisel kohal, kuid mõnes kõrvallauses viimasel kohal, olles samal ajal tundlik lause infostruktuuri suhtes (Lindström 2005, Tael 1988). Seega võivad sama tähendust kandva verbiühendi eri osad paikneda erinevas järjekorras (1).

(1) a. *Ta **jäi** kogemata **magama**.*
b. *Kui ta **magama jäi,** oli kell juba palju.*

Et vältida olukorda, kus *jäi magama* ja *magama jäi* loetakse eri tähendust kandvateks paarideks, järjestasin kõik kombinatsioonid infiniitse verbivormi kategooria järgi. Lõplik kandidaatandmestik näeb välja, nagu see on esitatud näites (2), kus esimeses tulbas on käändelise verbivormi kategooria kujul, nagu see on esitatud murdekorpuses, ning teises tulbas on sellega samas osalauses koosesinev pöördelise verbi algvorm.

(2) ma pidama
nud olema

inf maksma
ma pidama
ma pidama[4]

3. Terminoloogia

Enne konkreetselt meetodi kirjeldamise juurde asumist selgitan lühidalt kollostruktsioonilise analüüsi terminoloogiat. **Kolekseem** on sõna, mis on tugevalt seotud mõne konstruktsiooniga, seda nimetatakse selle konstruktsiooni kolekseemiks. **Kollostrukt** on konstruktsioon, mis on seotud ühe lekseemiga (sõnaga või mingi muu uuritava kategooriaga). **Kollostruktsioon** on seega kolekseemi ja kollostrukti kombinatsioon (Stefanowitsch, Gries 2003). Näiteks läänemurdes sage passiivikonstruktsioon *saama* ja *tud*-kesksõna on kollostruktsioon, kus kolekseem on *saama* ja kollostrukt seesama passiivikonstruktsioon (*tud* + *saama*).

Kollostruktsioonilised meetodid kasutavad kollokatsioonide tuvastamise meetoditest tuttavaid sõnadevahelise seose tugevuse mõõtmise statistikuid. Kandidaatandmestiku koostamisel järgitakse samu põhimõtteid. Korpusandmestiku põhjal koostatakse **kahemõõtmeline sagedustabel** (enamjaolt automaatselt), kus on esitatud sõnasagedus uuritavas konstruktsioonis, sama sõna sagedus kõigis teistes konstruktsioonides, konstruktsiooni sagedus, kus esinevad muud sõnad peale uuritava sõna (kolekseemi), kõigi teiste konstruktsioonide sagedus ilma uuritava sõnata (kolekseemita). Tabel 1 esitab kahemõõtmelise sagedustabeli jaoks vajalikud arvutused.

Tabel 1. Kahemõõtmeline sagedustabel.

O_{11}	O	$f_1 - O$	O_{12}
O_{21}	$f_2 - O$	N – f1 – f2 + O	O_{22}

4 Iga rida tähistab ühte ühendi esinemiskorda, st sagedusloendid ja vajalikud arvutused tehakse automaatselt programmiga, mida tutvustan osas 5.

Tabelis 2 on esitatud kahemõõtmeline sagedustabel *da*-infinitiivi ja finiitverbi *tahtma* paari kohta saarte murdes. Kokku oli selles murdes 3947 V_{fin} + inf kombinatsiooni (N=3947). Samas osalauses esinesid *tahtma* ja *-da* 18 korda (O=18). *da*-infinitiivi esines samas osalauses muu finiitverbiga 135 korda ning finiitne *tahtma* muu infiniitvormiga 16 korda. *da*-infinitiivi esines materjalis 153 korda ning *tahtma* 34 korda.

Tabel 2. Kahemõõtmeline sagedustabel (*da*-infinitiiv + *tahtma* saarte murdes).

		tahtma	**muu finiitverb**	
O_{11}	***da*-infinitiiv**	18	153 – 18 = 135	O_{12}
O_{21}	muu infiniitne vorm	34 – 18 = 16	3947 – 153 – 34 + 18 = 3778	O_{22}

4. Sõnadevahelise seose tugevuse mõõtmise statistikud

Nii kollokatsioonide kui ka konstruktsioonide analüüsimisel rakendatava statistiku väljavalimine on alati problemaatiline, kuna pole olemas ühte kindlat ja sobivaimat statistikut iga kindla ülesande jaoks. Selleks tuleb testida erinevaid statistikuid või toetuda eelnevalt tehtud uurimustele ja valida antud keele ja ülesande jaoks sobivaim[5].

Kollostruktsioonilise analüüsi jaoks on sobivaimaks peetud Fisher-Yatesi testi (*Fisher-Yates exact test*), kuna see statistik ei tee eeldusi andmestiku (valimi) suurusele ega jaotusele (Pedersen 1996, Stefanowitsch, Gries 2003). Seega on statistik eriti sobiv just keeleandmestiku jaoks, mille puhul normaaljaotus on ebatõenäoline. Samas on aga uurimusi, mis on näidanud mõne teise statistiku paremust Fisher-Yatesi testi (FYT) ees, näiteks minimaalse tundlikkuse mõõdik (*minimal sensitivity measure*) (Wiechmann 2008).

Murdekorpuse andmestik on selle poolest problemaatiline, et murded on materjali poolest erineva esindatusega. See

5 Statistikute kohta vt eestikeelset ülevaadet Uiboaed (2010).

omakorda tekitab olukorra, kus statistikute väärtused pole üldjuhul omavahel võrreldavad, kuna need on arvutatud erineva suurusega materjali põhjalt. Sellest asjaolust võib mööda vaadata juhul, kui soovitakse ainult näha, millised lekseemid tõmbuvad tugevamalt teatud konstruktsioonidega, ning ei soovitagi omavahel statistikute väärtusi võrrelda. Kui aga statistikute väärtusi peetakse tähenduslikuks, soovitakse neid võrrelda ning analüüsida, siis on materjali erinev suurus probleem.

On olemas statistikuid, mis esitavad omavahel võrreldavad väärtused ka erineva suurusega materjali põhjalt arvutatuna. Üks sellistest on riskisuhte mõõdik (*odds ratio*), mida kasutatakse just erineva suurusega korpustega töötades (Gries 2006). Käesolev artikkel jätkab siiski FYT-ga, kuna kasutan statistikuid siin vaid konstruktsioonide ekstraheerimiseks ja mitte nende väärtuste omavaheliseks võrdlemiseks. Murdekorpuse peal tehtud katsed osutavad selgelt konstruktsioonide tuvastamisel FYT-i paremusele muude statistikute ees (vt ka Uiboaed jt *ilmumas*). FYT arvutakse järgnevalt[6]:

$$\textit{Fisher-Yatesi test} = \sum_{k=O_{11}}^{\min(R_1 C_1)} \frac{\binom{C_1}{k} \cdot \binom{C_2}{R_1 - k}}{\binom{N}{R_1}}$$

Valemi jaoks vajalikud väärtused saadakse teises osas kirjeldatud kahemõõtmelise sagedustabeli põhjal ning *R* ja *C* osutavad vastavalt ridade ja veergude väärtustele.

6 Valem on esitatud nagu lehel http://www.collocations.de/AM/index.html (Evert 2004)

5. Kollostruktsioonilised meetodid

Kollostruktsioone on võimalik analüüsida kolmel erineval viisil. **Kolekseemide analüüsi** kasutatakse teatud konstruktsiooni ning selle konstruktsiooni kindlas positsioonis esinevate kolekseemide uurimiseks (Stefanowitsch, Gries 2003). Näiteks võib uurida *ma*-infinitiivis esinevat verbi kausatiivses konstruktsioonis, kus pöördeline vorm on verb *panema* (*pani nutma*), st millised *ma*-infinitiivis esinevad verbid tõmbuvad mõne *panema* pöördelise vormiga. See meetod on edasiarendus lihtsatest kollokatsioonide leidmise meetoditest (Gries, Stefanowitsch 2010).

Eristavate kolekseemide analüüs (Gries, Stefanowitsch 2004) uurib sünonüümseid konstruktsioone ning nendega seotud sõnu, näiteks kausatiivset konstruktsiooni pöördelise verbiga *panema*, kus teine osa on kas verb *ma*-infinitiivis või hoopis mõni nimisõna (*paneb põlema vs paneb tööle*). Meetod võimaldab analüüsida, kas konstruktsiooni nimisõnalise või verbilise osise semantikas on mingeid erinevusi, kas mingite tähenduste väljendamiseks kasutatakse selles konstruktsioonis pigem nimisõnu või *ma*-infinitiivis verbi. Konstruktsiooni tähenduse üle saab peale statistiliste analüüside läbiviimist otsustada keeleuurija, st konstruktsiooni tähendus ei ilmne kusagilt automaatselt.

Koosvarieeruvate kolekseemide analüüs uurib ühe konstruktsiooni erinevates positsioonides esinevate sõnade omavahelisi suhteid (Stefanowitsch, Gries 2005), näiteks millised verbid esinevad potentsiaalsete verbikonstruktsioonide infiniitses ja millised finiitses positsioonis ja kui tugev on nende omavaheline seos. See meetod on sarnane traditsiooniliste kollokatsioonide tuvastamise meetoditega selle erinevusega, et tähelepanu suunatakse ainult kahe keeleüksuse semantilistele ja/või süntaktilistele omadustele.

Artikli järgnev osa keskendub vaid viimasele kirjeldatud meetoditest, koosvarieeruvate kolekseemide analüüsile. Uurin verbi morfoloogilise kategooria ja sellega koosesineva finiitverbi omavahelisi suhteid. Keskendun infiniitse verbi morfoloogilisele kategooriale ning temaga koosesinevale finiitse verbi lemmale. Selle meetodiga peaksid esile kerkima ühendid, milles finiitverb on suure tõenäosusega mingil määral grammatiseerunud.

Statistikute väärtuste arvutamiseks on käesolevas artiklis kasutatud Coll.analysis.3 programmi (Gries 2007), mis on välja töötatud just kollostruktsioonilise analüüsi tarbeks ning on vabalt kasutatav. Programm on hästi dokumenteeritud ja statistikaprogrammiga R (R Development Core Team 2011) kergesti rakendatav.

Niisiis keskendub artikkel infiniitse verbivormi morfoloogilise kategooria ning finiitse verbi ühendite koosvarieeruvate kolekseemide analüüsile eesti murretes. Vaatlen, millised vormid ja verbid paiknevad selliste ühendite erinevates positsioonides ning millised neist on piisavalt kõrge statistiku väärtusega, et neid 95% tõenäosusega konstruktsioonideks pidada. Uurin, kas eri murretes on selles osas mingit varieerumist. Kui jah, siis milliste konstruktsioonide osas ja millised murdegrupid sellistest erinevustest joonistuvad.

Selleks, et mõõta statistilist seost kahe uuritava keeleüksuse vahel programmiga Coll.analysis.3, on vajalik eelnevalt korpusest väljavõetud paaridega fail iga murde kohta eraldi kujul, mis on esitatud eespool näites (2), kus iga finiitverbi ja infiniitse vormi esinemiskord on eraldi real. Esimeses tulbas on infiniitse verbivormi märgend murdekorpuses esitatud kujul, ning teises tulbas selle vormiga samas osalauses koos esinev finiitse verbi lemma. Programm teeb arvutused kõigi leitud paaride põhjal, kasutades infot, mis on esitatud eespool kirjeldatud kahemõõtmelise sagedustabelina.

Coll.analysis.3. programmi suur eelis on, et kogu arvutuskäigu viib läbi programm ning keeleuurijal tuleb valida ainult endale sobiv statistik ja tulemuste esitamisviis.

6. Tulemused

6.1. Kollostruktsioonilise analüüsi tulemused

Kui iga murde jaoks on arvutatud seal esinevate verbiühendite seose tugevuse mõõdiku väärtused, võib neid edasi analüüsida. Jätkan ainult nende ühenditega, mille kohta võib 95% tõenäosusega väita, et tegemist on konstruktsioonide ja

mitte juhuslike koosesinemistega. Edaspidi nimetangi statistiliselt olulisi ühendeid konstruktsioonideks. Jätan analüüsist välja ühendid, mis esinesid korpuses vähem kui kolm korda ja/või mille statistiline seos oli nõrk.

Tabel 3 esitab kõik kollostruktsioonilise analüüsi tulemusel tuvastatud konstruktsioonid iga murde kohta eraldi. Tabelis ei ole esitatud statistiku väärtusi, kuna selles töös ei oma need tähtsust, samuti pole need murrete lõikes võrreldavad, kuna kasutasin FYT-i (vt artikli 4. osa). Niisiis on tabelis esitatud tähestikuliselt iga murde verbikonstruktsioonid, mis kollostruktsioonilise analüüsi käigus tuvastati.

Kõige rohkem on tuvastatud erinevaid konstruktsioone saarte, lääne- ja keskmurdes, kõige vähem Seto, Mulgi ja kirdemurdes. Selliseid konstruktsioone, mis kõigis murretes tuvastati, on vaid neli (*ma*-infinitiiv + *hakkama, minema, panema, pidama* ning *nud*-partitsiip + *olema*). Võib öelda, et esitatud finiitverbid on kõigis murretes üsna tugevalt grammatiseerunud ning moodustavad kindlalt vastava infiniitvormiga koosesinedes konstruktsiooni. Teised kõigis murretes, v.a Seto, tuvastatud konstruktsioonid on *mas*-vormi + *käima* ning *tud*-partitsiibi + *olema* konstruktsioonid.

Tabelis võib paraku näha ka süntaksianalüsaatori vigadest tingitult moodustunud V_{fin} + inf paare, kui osalause piire ei tuvastata korrektselt, näiteks on *tud*-partitsiip ja *katkuma* saarte murdes üks selliselt moodustunud paare, mis on saanud kõrge FYT-i väärtuse. Kvalitatiivse analüüsi käigus selgus, et tegemist on osalausestamisveaga. Sellised vead on automaatsel keeletöötlusel vältimatud, eriti automaattöödeldavuselt nii probleemse materjali puhul, kui seda on murdematerjal. Seetõttu annab parima tulemuse kindlasti kvalitatiivsete ja kvantitatiivsete meetodite kombineerimine, mida murdesüntaksi uurimisel siiani pole eriti rakendatud.

Kuna sellistest pikkadest konstruktsiooniloeteludest nagu tabelis 3 on palja silmaga raske mingit terviklikumat ülevaadet saada, tuleb andmeid kuidagi visualiseerida. Järgnev osa kirjeldabki seda, kuidas kollostruktsioonilise analüüsi tulemuste abil murdeid verbikonstruktsioonide ja nende sageduste põhjal grupeerida ja tulemusi visualiseerida.

Tabel 3. Tuvastatud kollostruktsioonid FYT-ga 95% usaldusnivool. Lühendid: inf – *da*-infinitiiv, ma – *ma*-infinitiiv, mas – *ma*-infinitiivi inessiiv, mast – *ma*-infinitiivi elatiiv, mata – *ma*-infinitiivi abessiiv, nud – *nud*-partitsiip, tud – *tud*-partitsiip, *v* – *v*-partitsiip.

IDA	KESK	KIRDE	LÄÄNE	MULGI	RANNA	SAARTE	SETO	TARTU	VÕRU
inf_jõudma	inf_andma	inf_saama	inf_laskma	inf_tulema	inf_saama	inf_käskima	inf_saama	inf_jõudma	inf_saama
inf_laskma	inf_julgema	ma_hakkama	inf_oskama	ma_hakkama	inf_tahtma	inf_laskma	ma_hakkama	inf_saama	inf_tulema
inf_saama	inf_laskma	ma_jääma	inf_saama	ma_minema	inf_tulema	inf_lubama	ma_minema	ma_hakkama	ma_hakkama
inf_tahtma	inf_tahtma	ma_minema	inf_tahtma	ma_panema	inf_täidima	inf_oskama	ma_panema	ma_heitma	ma_minema
inf_võima	inf_tohtima	ma_panema	inf_tulema	ma_pidama	ma_hakkama	inf_saama	ma_pidama	ma_jääma	ma_panema
ma_hakkama	inf_tulema	ma_pidama	inf_võima	mas_käima	ma_jääma	inf_tahtma	ma_tulema	ma_minema	ma_pidama
ma_minema	inf_võima	mas_käima	ma_ajama	nud_olema	ma_minema	inf_tohtima	nud_olema	ma_panema	ma_tulema
ma_panema	ma_ajama	mast_tulema	ma_hakkama	nud_tegema	ma_panema	inf_tulema	tud_olema	ma_pidama	mas_käima
ma_pidama	ma_hakkama	nud_olema	ma_juhtuma	nud_vaatama	ma_pidama	inf_viitsima	v_ajama	ma_tulema	mas_olema
ma_tulema	ma_juhtuma	nud_teadma	ma_jääma	tud_olema	ma_tulema	inf_võima		mas_käima	mata_jääma
mas_käima	ma_jääma	tud_olema	ma_kippuma	tud_saama	mas_käima	ma_ajama		mata_jääma	nud_kaema
mata_jääma	ma_minema	tud_saama	ma_minema		mata_jääma	ma_hakkama		nud_olema	nud_olema
nud_olema	ma_panema		ma_panema		nud_olema	ma_juhtuma		tud_olema	tud_olema
nud_rääkima	ma_pidama		ma_pidama		nud_surema	ma_jääma		tud_saama	tud_saama

tud_olema	ma_tulema		ma_õpetama		nud_taidma	ma_kukkuma	
tud_saama	ma_õppima		mas_käima		nud_vaatama	ma_minema	
tud_teadma	mas_käima		mata_jäämа		tud_saama	ma_panema	
	nud_magama		nud_kustuma			ma_pidama	
	nud_naerma		nud_kuulma			ma_saatma	
	nud_olema		nud_olema			ma_tulema	
	nud_rääkima		nud_rääkima			mas_käima	
	nud_tegema		nud_teadma			mata_jäämа	
	nud_tooma		nud_vaatama			nud_magama	SAARTE
	nud_vaatama		nud_ütlema			nud_olema	tud_katkuma
	nud_valama		tud_saama			nud_rääkima	tud_saama
	nud_võtma		tud_tegema			tud_heitma	tud_tegema
	tud_saama		v_panema			tud_hoidma	v_panema

6.2. Murrete grupeerimine kollostruktsioonilise analüüsi ja konstruktsioonide sageduste põhjal

Eelnevalt kirjeldasin kollostruktsioonilist analüüsi ja seda, kuidas statistiliselt keelekonstruktsioonid murdekorpusest tuvastatud on. Tabelis 3 on loetelu kõigist tuvastatud konstruktsioonidest. Huvitav oleks aga teada, kuidas need konstruktsioonid murrete lõikes varieeruvad. Milliseid gruppe murded nende põhjal moodustavad ning millised konstruktsioonid on mingile murdele iseloomulikumad?

Neile küsimustele vastamiseks kasutan üht mitmemõõtmelise statistika meetodit – korrespondentsanalüüsi. Käesoleva artikli maht ei võimalda laskuda meetodi arvutuslikesse detailidesse, vaid keskendub ainult lõpptulemuse selgitamisele[7]. Korrespondentsanalüüsi kasutatakse mitmemõõtmelisest sagedusandmestikust varjatud mustrite tuvastamiseks. Meetod on pigem esmane vahend andmestikust ülevaate saamiseks ning seda kasutatakse sageli rohkem edasise analüüsi jaoks hüpoteeside püstitamiseks kui põhjapanevate statistiliste järelduste tegemiseks.

Korrespondentsanalüüsi sisendiks on tavaline sagedustabel, kus ridades on konstruktsioonid ja nende sagedused ning veergudes murded. Siin on oluline tähele panna, et nüüd ei võrdle ma enam statistikute väärtusi, vaid kaasan analüüsi ainult piisavalt kõrge statistiku väärtusega ühendid ehk konstruktsioonid (vt osa 6.1). Kuna eelmise analüüsi tulemusena tuvastasin kõik konstruktsioonid, võin edasi võrrelda vaid tuvastatud konstruktsioonide normaliseeritud sagedusi[8] ja nii nagu eelnevaski analüüsis, on ka siin eemaldatud madala esinemissagedusega ühendid. Korrespondentsanalüüs võtab sisendiks mitmemõõtmelise sagedustabeli ning esitab kogu tabelis sisalduva informatsiooni

7 Korrespondentsanalüüsi kohta täpsemalt vt näiteks Greenacre (2007), Baayen (2008:139–146) ja murdekorpuse kontekstis Uiboaed jt (*ilmumas*).

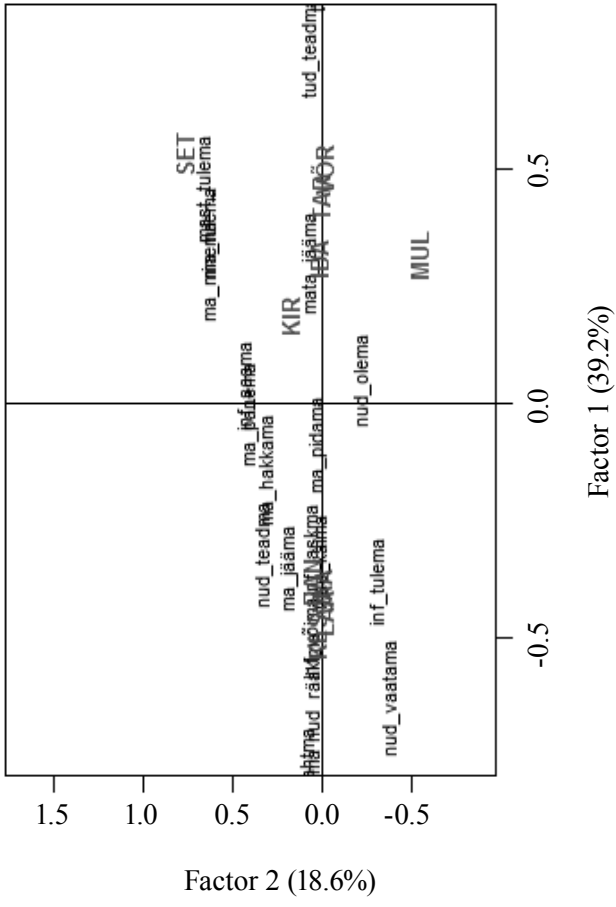
8 Sagedused on normaliseeritud, et neid omavahel võrrelda. Kuna murdekorpuses pole kõik murded materjali poolest võrdselt esindatud, pole lihtsagedused omavahel võrreldavad. Normaliseerimine on tehtud keskmise korpuse suuruse põhjal.

kahemõõtmelisena. Joonisel 2 on esitatud korrespondentsanalüüsi tulemused.[9] Suurte ja heledamate tähtedega on tähistatud murded ning väikeste ja tumedatega konstruktsioonid. Esimese, kõige olulisema dimensiooni (horisontaalse) põhjal on murrete varieeruvus märkimisväärne ning moodustub kaks väga selget rühma. Esimese moodustavad kesk-, lääne-, ranna- ja saarte murre ning teise kirde-, ida-, Mulgi, Seto, Tartu ja Võru murre. Esimene grupp näib olevat üsna homogeenne, kõik murded paiknevad tihedalt üksteise lähedal, st on tugevasti konstruktsioonide ja nende sageduste põhjal seotud. Teine grupp eristub selgelt esimesest, kuigi grupisisene varieeruvus on suurem.

Teise dimensiooni (vertikaalse) põhjal erinevad muudest murretest kõige rohkem Seto ja Mulgi, mis on ka omavahel üsna erinevad. Niisiis kõik murded, mis graafikul üksteise lähedal paiknevad, on omavahel konstruktsioonide ja nende sageduste põhjal sarnasemad, ning need, mis üksteisest kaugemal asetsevad, erinevamad. Sama kehtib konstruktsioonide kohta. Mida lähemal nad mingile murdele graafikul paiknevad, seda iseloomulikumad nad sellele murdele on, ja vastupidi.

Korrespondentsanalüüsi tulemusena ilmnevad huvitavad tendentsid. Graafikult torkab kohe silma, et esitatud jaotus ei järgi traditsioonilist murdejaotust, kus tugevaim erinevus on põhja- ja lõunapoolsete murrete vahel. Selles töös uuritud nähtuse põhjal on erinevus suurim ida- ja läänepoolsete murrete vahel. Läänepoolse rühma moodustavad kesk-, lääne-, saarte ja rannamurre ning ülejäänud murded moodustavad selgelt teise, idapoolsema rühma. Konstruktsioonide kasutuses torkab silma suurem konstruktsioonide arv, mis on tugevalt seotud läänerühmaga, samas kui idarühmaga on tugevalt seotud vähesed konstruktsioonid. Idarühmale on iseloomulikumad näiteks *ma*-infinitiivi + *minema, tulema* konstruktsioonid ja *mast*-vormi + *tulema* konstruktsioon. Viimane on eriti iseloomulik Seto murdele. *mata*-vormi + *jääma* konstruktsioon on tihedalt seotud idamurdega. Mõlemas rühmas esinevad suhteliselt ühtlaselt näiteks

9 Tulemused esitavad kaks esimest dimensiooni. Esimene dimensioon (horisontaalne) seletab 39,2% andmestiku variatiivsusest ning teine (vertikaalne) 18,6%.



Joonis 2. Korrespondentsanalüüsi tulemused. Lühendid: RAN – ranna, IDA – ida, SAA – saarte, KES – kesk, MUL – Mulgi, KIR – kirde, SET– Seto, TAR – Tartu, LÄÄ – lääne, VÕR – Võru.

ma-infinitiiv + *hakkama, minema, panema, pidama* ning *nud*-partitsiip + *olema* konstruktsioonid (paiknevad joonisel pigem kahe rühma keskel). Eelmises osas tuli välja, et just need konstruktsioonid tuvastati kõigis murretes, seega pole nad ühelegi murdele eriliselt iseloomulikud, vaid esinevad kõigis murretes üsna homogeenselt. Läänepoolsele rühmale on iseloomulikud näiteks *da*-infitiivi ja *võima, laskma, tulema* konstruktsioonid ja *ma*-infitiivi + *jääma* ja *mas*-vormi + *käima* konstruktsioonid.

Kokkuvõtteks võib öelda, et verbikonstruktsioonide kasutus eesti murretes ei ole seotud traditsioonilise murdejaotusega, vaid rühmad moodustuvad selgelt ida- ja läänepoolsete murrete erinevuste põhjal. Edasi peaks kindlasti uurima, kuidas jagunevad nende konstruktsioonide tähendused murretes ning kas sealt tulevad välja samasugused erinevused, kuid see jääb juba edasise uurimistöö teemaks.

7. Kokkuvõte

Artiklis tutvustasin ühte kvantitatiivse korpuslingvistika meetoditest – kollostruktsioonilist analüüsi, rakendades seda eesti murrete korpuse materjali peal. Artikkel uuris verbi finiitse ja infiniitse vormi konstruktsioonide varieerumist eesti murretes, täpsemalt verbi infiniitse vormi morfoloogilise kategooria ja finiitse verbi lemma konstruktsioone. Tutvustasin analüüsiks kasutatavat terminoloogiat ja materjali ning selle saamise viisi. Iga murde kõigi potentsiaalsete konstruktsioonide jaoks arvutasin Fisher-Yatesi testi väärtused ning edasi analüüsisin vaid ühendeid, mille kohta võis piisava statistilise kindlusega väita, et tegemist on konstruktsioonidega, mitte uuritavate paaride osiste juhuslike koosesinemistega. Samuti jäid analüüsist välja ühendid, mis esinesid korpuses vähem kui kolm korda. Lõplikus analüüsis kasutasin seega statistiliselt tuvastatud verbikonstruktsioone ja nende normaliseeritud sagedusi.

Edasi kasutasin korrespondentsanalüüsi, et tuvastada andmestikus olevaid varjatud mustreid, mis võiks seletada murrete varieerumist konstruktsioonide ja nende normaliseeritud

sageduste põhjal. Tulemused osutavad, et murded ei grupeeru uuritavate konstruktsioonide põhjal nii nagu traditsiooniline murdejaotus. Verbikonstruktsioonide kontekstis on suuremad erinevused ida- ja läänepoolsete murrete vahel ning traditsioonilise põhja-lõuna grupi erinevused siit eriti tugevalt ei ilmne. Läänepoolsetele murretele on erinevate verbiühendite kasutamine iseloomulikum kui idapoolsetele murretele, st analüüs tuvastas, et erinevad konstruktsioonid on tugevamalt seotud läänerühmaga ning idarühmaga tõmbuvad tugevamalt vaid üksikud verbikonstruktsioonid. Tuvastatud konstruktsioonide tähendusnüansside uurimine jääb edasise uurimistöö teemaks.

Aadress:
Kristel Uiboaed
Eesti ja üldkeeleteaduse instituut
Tartu Ülikool
Jakobi 2, 51014 Tartu
E-mail: Kristel.Uiboaed@ut.ee

Kirjandus

Baayen, R. Harald (2008) *Analyzing linguistic data: a practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Erelt, Mati, Reet Kasik, Helle Metslang, Henno Rajandi, Kristiina Ross, Henn Saari, Kaja Tael, Silvi Vare (1993) *Eesti keele grammatika II. Süntaks. Lisa: kiri*. Tallinn: Eesti Teaduste Akadeemia Keele ja Kirjanduse Instituut.

Evert, Stefan (2004). *Computational approaches to collocations*. <http://www.collocations.de/>. Vaadatud 15.10.2012.

Evert, Stefan (2005) *The statistics of word cooccurrences: word pairs and collocations*. Universität Stuttgart. Available online at <http://www.bsz-bw.de/cgi-bin/xvms.cgi?SWB12046165>. Vaadatud 15.10.2012.

Evert, Stefan (2008) „Corpora and collocations". In M. Kytö and A. Lüdeling, eds. *Corpus linguistics: an international handbook*. Berlin: Mouton de Gruyter.

Gilquin, Gaëtanelle (2006) The verb slot in causative constructions: finding the best fit. *constructions*. (Supplement, 1). Osnabrück, Germany. Available online at <http://search.ebscohost.com/login.aspx?direct=true&db=mzh&AN=2008900938&site=ehost-live>. Vaadatud 15.10.2012.

Goldberg, Adele E. (1995) *Constructions: a construction grammar approach to argument structure*. Chicago: University of Chicago Press.

Greenacre, Michael (2007) *Correspondence analysis in practice*. 2nd ed. London, New York: Chapman & Hall, CRC Press.

Gries, Stefan Th. (2006) "Exploring Variability within and between corpora: some methodological considerations." *Corpora: Corpus-Based Language Learning, Language Processing and Linguistics* 1, 2, 109–151.

Gries, Stefan Th. (2007) Coll.analysis 3.2. A program for R for Windows 2.x.

Gries, Stefan Th. and Anatol Stefanowitsch (2004) "Extending collostructional analysis: a corpus-based perspective on "alternations"". *International Journal of Corpus Linguistics* 9, 1, 97–129.

Gries, Stefan Th. and Anatol Stefanowitsch (2010) "Cluster analysis and the identification of collexeme classes". In Sally Rice and John Newman, eds. *Empirical and experimental methods in cognitive/functional research*, 73–90. Stanford, CA: CSLI.

Hilpert, Martin (2006) "Distinctive collexeme analysis and diachrony". *Corpus Linguistics and Linguistic Theory*, 2(2), 243–256.

Hilpert, Martin (2008) *Germanic future constructions: a usage-based approach to language change*. Amsterdam and Philadelphia: John Benjamins.

Lindström, Liina (2005) *Finiitverbi asend lauses. Sõnajärg ja seda mõjutavad tegurid suulises eesti keeles.* (Dissertationes philologiae Estonicae Universitatis Tartuensis.) Tartu: Tartu Ülikooli Kirjastus.

Lindström, Liina, Liisi Bakhoff, Mari-Liis Kalvik, Anneliis Klaus, Rutt Läänemets, Mari Mets, Ellen Niit, Karl Pajusalu, Pire Teras, Kristel Uiboaed, Ann Veismann, Eva Velsker (2006) "Sõnaliigituse küsimusi eesti murrete korpuse põhjal". Niit, Ellen, toim. *Keele ehe*, 154-167. Tartu: Tartu Ülikooli eesti keele õppetool.

Lindström, Liina ja Kaili Müürisep (2009) "Parsing Corpus of Estonian Dialects." *Proceedings of the NODALIDA 2009 workshop Constraint Grammar and robust parsing*, 22–29. Northern European Association for Language Technology.

Mukherjee, Joybrato and Stefan Th. Gries (2009) "Collostructional nativisation in New Englishes: Verb-construction associations in the International Corpus of English". *English World-Wide* 30, 1, 27–51.

Müürisep, Kaili (2000) *Eesti keele arvutigrammatika: süntaks.* (Dissertationes Mathematicae Universitatis Tartuensis.) Tartu: Tartu Ülikooli Kirjastus.

Pedersen, Ted (1996) "Fishing for Exactness." In *Proceedings of the South Central SAS User's Group Conference (SCSUG-96),* Austin, TX, Oct 27-29, 1996, 188–200.

Penjam, Pille (2008) *Eesti kirjakeele* da- *ja* ma-*infinitiiviga konstruktsioonid*. (Dissertationes philologiae Estonicae Universitatis Tartuensis.) Tartu: Tartu Ülikooli Kirjastus.

R Development CoreTeam (2011) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.

Sahkai, Heete (2011) *Teine grammatika. Eesti keele teonimede süntaks konstruktsioonipõhises perspektiivis*. (Tallinna Ülikooli humanitaarteaduste dissertatsioonid.) Tallinn: Tallinna Ülikooli Kirjastus.

Stefanowitsch, Anatol and Stefan Th. Gries (2003) "Collostructions: Investigating the interaction of words and constructions". *International Journal of Corpus Linguistics* 8, 2, 209–243.

Stefanowitsch, Anatol and Stefan Th. Gries (2005) "Covarying collexemes". *Corpus Linguistics and Linguistic Theory* 1, 1, 1–43.

Stefanowitsch, Anatol and Stefan Th. Gries (2009) "Corpora and grammar". Anke Lüdeling and Merja Kytö, eds. *Corpus linguistics: an international handbook*. Vol. 2, 933-951. Berlin, New York: Mouton de Gruyter.

Tael, Kaja (1988) „Sõnajärjemallid eesti keeles (võrrelduna soome keelega)." *Läänemeresoome keeleteaduse sümpoosion, Turku, 30.08-2.09. 1988*. (Preprint KKI-56.) Tallinn: Eesti NSV Teaduste Akadeemia Keele ja Kirjanduse Instituut.

Uiboaed, Kristel, Cornelius Hasselblatt, Liina Lindström, Kadri Muischnek, John Nerbonne (ilmumas) "Constructional variation in Estonian dialects." *LLC: The Journal of Digital Scholarship in the Humanities*.

Uiboaed, Kristel (2010) „Statistilised meetodid murdekorpuse ühendverbide tuvastamisel." *Eesti Rakenduslingvistika Ühingu Aastaraamat* 6, 307–326.

Wiechmann (2008) "On the computation of collostruction strength: testing measures of association as expressions of lexical bias". *Corpus Linguistics and Linguistic Theory* 4, 2, 253–290.

Abstract. Kristel Uiboaed: Collostructional methods and verb constructions in Estonian dialects. The present work studies finite and non-finite verb constructions and their variation in Estonian dialects. Article gives an overview of collostructional methods and applies the method to detect verb constructions in Estonian dialects. Detected constructions are studied further to explore which finite verbs show grammaticalization tendencies in different dialects. Construction-based division of dialects is presented. Results indicate that construction based classification of dialects leads to different groups compared to traditional dialect classifications. Major differences occur between eastern and western dialects, whereas western dialects use different verb constructions considerably more than eastern dialects do.

Keywords: collostructional methods, chain verbs, corpus linguistics, dialectology, dialect syntax, verb constructions